Thaisen, J. (2013) Gamelyn's Place among the Early Exemplars for Chaucer's Canterbury Tales . *Neophilologus*, 97(2), pp. 395–415

Link to official article DOI :10.1007/s11061-012-9315-3
(Access to content may be restricted)

Gamelyn's Place among the Early Exemplars for Chaucer's Canterbury Tales

Jacob Thaisen

Abstract
Application of standard techniques from natural language processing on N-gram models of spelling enables quantification of the similarity between Middle English texts despite their lexical differences. Three studies employing similarity metrics confirm that a scribe's spelling always is biased in the direction of his exemplars. This bias opens up a window on the number of scribes behind the exemplars for a text executed in a single hand, when other variables such as authorship and poetic form are held constant. A fourth study addresses nine manuscripts of Geoffrey Chaucer's poem the Canterbury Tales with early textual contents. A one-way ANOVA/Tukey's Range Test shows that none of these manuscripts is based on exemplars written in more than three hands, when allowance is made for variation due to poetic form. The results point to unified exemplars for the full text as the normal format for the poem's transmission. The discussion suggests that the final tale ordering found in the first manuscripts is a product of collaboration between the poem's first two scribes, probably working after Chaucer's death and spuriously adding the Tale of Gamelyn.
Keywords Scribal copying · Natural language processing · Authorship attribution · Chaucer · Canterbury Tales · Manuscript production

Did the poet Geoffrey Chaucer compose the Tale of Gamelyn? The question is of especial interest in view of two recent developments. Firstly, there is increasing recognition today that Chaucer may have overseen the initial phase of the scribal

To Norman F. Blake.

work on the Hengwrt manuscript of his Canterbury Tales (Aberystwyth, National Library of Wales, Peniarth 392D), and even that this count of authorially supervised manuscripts of the poem may be a conservative one. The emerging picture is of piecemeal compilation of the first manuscripts as a result of the poet continuing to add to, revise, and rearrange his text up until his premature death. Their scribes, further, have associations with the London Guildhall, which make it possible to imagine them exchanging exemplars for the poem with one another. Secondly, a trace of the history of the exclusion of the Tale of Gamelyn from the Chaucer canon has revealed its exclusion primarily to rest on perceived deficiency in its poetic quality, rather than any strong evidential basis (Vázquez 2009). These recent developments make it a realistic possibility that the poet himself discussed the possible inclusion of the Tale of Gamelyn with the first scribes. They also make it conceivable that he is its author.

This paper argues against any such Chaucerian sanctioning of the Tale of Gamelyn. It first discusses three studies to establish similarity metrics based on N-gram models of spelling as a discriminator of scribal hands. It next identifies the number of scribal hands behind the respective exemplars for each of nine manuscripts of the Canterbury Tales with an early text, before exploring the implications of this number for the textual tradition of the poem. The number would seem to be no more than three for any individual manuscript, distributed such that the exemplars for the Tale of Gamelyn never share their hand with any canonical text. Their separateness would appear to confirm that the tale is a spurious addition to the poem. The metrics may further suggest that the text of the poem soon came normally to be transmitted as a whole, two scribes possibly having prepared exemplars of their own from Chaucer's draft.

Previous studies of spelling have failed to dismiss possible Chaucerian authorship of the Tale of Gamelyn. The tale has been found to have a spelling profile very similar to that of the canonical tales surrounding it for both the Oxford, Corpus Christi College 198 and London, British Library, Harley 7334 manuscripts, but not to do so for manuscript Oxford, Christ Church 152 (Blake and Thaisen 2004; Thaisen 2008a). The compilation of spelling profiles for these three manuscripts followed the general procedure described in McIntosh et al.'s (1986) A Linguistic Atlas of Late Medieval English, and the availability of electronic transcripts prepared by the Canterbury Tales Project made it feasible to include numerous spelling forms and map out their occurrence in the entire text of the manuscripts. The analysis paid attention to differences in how spelling forms are distributed between sections of the manuscripts, only subsidiarily to the forms' quality at other levels of language.

However, for the purpose of establishing and analysing how spelling forms are distributed in a manuscript, tabular arrangement of them and their frequencies of occurrence in individual sections may mislead, irrespective of how many forms are included. The difficulty to identify an anomalous section through visual analysis arises not only because sectioning may blur an anomaly by accidentally failing to coincide with it, but also because an anomalous section simply may fail to stand out. The former issue can be easily remedied by comparing results obtained with two or

more separate divisions of a text, but the non-categorical nature of variations in spelling make the latter issue harder to safeguard against. An anomaly typically manifests itself as an aggregate difference in the relative occurrence of individual spelling forms, which means that no pattern may be readily apparent at the level of the individual form. The constraints of a table's two-dimensional structure similarly means that it juxtaposes certain spelling forms, thus possibly preventing the researcher from detecting correlations in how other forms are distributed. Reasons of this kind may explain the apparent integration, at the level of spelling, of the Tale of Gamelyn with the surrounding canonical tales in Corpus 198 and Harley 7334. Specifically, Blake and Thaisen (2004) and Thaisen (2008a) saw a gradual change in the relative use of spelling forms where the present metrics strongly indicate an abrupt change.

Probabilistic models

What constitutes a more accurate basis for comparison is an exhaustive inventory of the building blocks of all spelling forms and their frequencies of occurrence, particularly an inventory statistically optimised with respect to its salient features. A smoothed, interpolated N-gram model has these properties. The present study relies on such models.[1]

The fundamental idea is to use N-gram models of spelling to quantify similarity between texts. Such a model is substantially identical to a traditional linguistic profile enumerating spelling forms and their frequencies but can be considered more objective. It is straightforwardly a full inventory of the letter sequences of length N that occur in a training text and their frequencies. Every such sequence, or N-gram, found in a test text receives a separate probability established from its frequency according to the model plus a weighting. The term 'smoothing' refers to a family of techniques, routinely used in natural language processing, for calculating weights so as to reduce a model's dependence on the training text's lexicon, in effect for transforming a distribution with statistical outliers into a more normal distribution. Together with 'interpolation', smoothing enables a model to assign probability to spelling forms that are predictable but unattested. An interpolated N-gram model is one incorporating a separate model for every gram length shorter than N. To assign probability to an N-gram absent from an N-gram model, it is dissolved into its two constituent (N-1)-grams, which may be attested. A metric derived from the average of the probabilities numerically expresses the level of similarity between a model and a test text. This metric is called perplexity, a low perplexity indicating great similarity.

[1] The most promising alternatives adopt Burrow's 'Delta' as a model, which is a stylometric test based on differences in items' ranking on frequency lists. It has been shown that the most frequent items discriminate well in lexically-based comparisons of texts written in analytical languages. For synthetic languages, however, discrimination improves with items taken from regions at various distances from the top of the ranked frequency list. This finding invites further research to identify what area of the frequency list discriminates best for comparisons based on Middle English spelling forms. See Rybicki and Eder (2011) for discussion and further references.

To the best of my knowledge, little precedent exists for adopting perplexity of N-gram models as a similarity metric with medieval English spelling data. The metric has been used in three studies, of which two are published. The first of them confirmed N-gram models of spelling as a suitable basis for comparison of lexically different texts (Thaisen 2009). The variables authorship and poetic form were kept constant, as both the training and test data consisted of two tales from the Canterbury Tales written in iambic pentameter end-rhymed verse. Specifically, the data were every medieval copy of the Miller's Tale and every such copy of the Wife of Bath's Prologue, totalling 116 texts. The transcripts were published (Robinson 1996, 2004), and accorded with the transcription conventions described in Robinson and Solopova (1993). The modelling software was the SRI Language Modelling Toolkit ('SRILM') freely available from SRI International (Stolcke 2002).[2] The individual text served as training data for a corresponding model, which in all cases were interpolated 3-gram ones smoothed according to the method developed by Witten and Bell (1991). All punctuation having been removed and all emphatic letter shapes having been made nonemphatic (lowercase), a spelling form was defined as a sequence of one or more letters occurring between two spaces. The training and test data both included grams spanning the space separating consecutive spelling forms, but excluded grams containing a line-boundary marker. Every model was tested on every text. The resulting perplexities were hierarchically clustered and the clusters visualised by means of two dendrograms: one for models based on the Miller's Tale and the other for models based on the Wife of Bath's Prologue.

The two dendrograms were essentially mirror images of one another. In both dendrograms, the terminal nodes representing the copy of the Miller's Tale from a given manuscript and the copy of the Wife of Bath's Prologue from the same manuscript were found practically always to be siblings. Only in a handful of cases did the same mother not dominate the nodes corresponding to the two copies taken from any one manuscript. In those few cases, the two copies were known on paleographical evidence to be by different scribes or from textual scholarship to be stemmatically distant; the latter may suggest exemplars by different scribes at a prior stage in the transmission of the text. General similarity of spelling unified the nodes dominating these mothers. The manuscripts thought by paleographers to share their scribe, such as London, British Library, MSS Additional 5140 and Egerton 2864, thus clustered together, as did those characterised by the presence of western features, those in the 'Devonshire group' recognisable by their hooked g (Mooney and Mosser 2004),[3] and those associated with Type III London English and the closely related tale orders of San Marino, California, Huntington Library, MS El.26.C.9 ('Ellesmere') and the a manuscripts. Another node contained William

[2] http://www-speech.sri.com/projects/srilm/. Accessed 25 December 2011.

[3] The group derives its name from the 'Devonshire' manuscript, today Tokyo, Professor Takamiya of Keio University, 24. The text of this manuscript is affiliated with London, British Library, Egerton 2726 and Dd.4.24. The similarity metrics, however, has it clustering with its linguistic and paleographical peers, Cambridge, Trinity College R.3.3 and Oxford, Bodleian Library, Rawlinson poet. 223.

Caxton's two print editions; the printer famously revised his earlier edition by consulting a separate exemplar to produce the later edition. The members of yet another node were Richard Pynson's and Wynkyn de Worde's print editions of the poem; Pynson set his edition from de Worde's.

Could it be concluded that the similarity of spelling evident between the respective copies of the Miller's Tale and the Wife of Bath's Prologue from any one manuscript arose exclusively because they share a scribe? The textual and codicological evidence has been read to suggest a series of separate exemplars, each with its own history, as the regular format in which the text of the Canterbury Tales was transmitted throughout much of the fifteenth century (Manly and Rickert 1940; Owen 1991). It is reasonable to expect separate exemplars to have led to differences in spelling, unless every scribe thoroughly converted what he found in his exemplars into his own kind of spelling. As scribes are known routinely to carry out partial conversions rather than complete ones (McIntosh 1963; Benskin and Laing 1981), the invisibility of any such separate exemplars from the dendrograms rather suggested the more nuanced alternative that similarity of spelling was in evidence already between the exemplars drawn on by the individual scribe, which implies that the poem was normally transmitted as a unified whole. The text may have consisted of a series of physically discrete exemplars, but if so, the default was for those exemplars to be in a single scribal hand and to travel together. The analysis presented in Manly and Rickert (1940) can be read to support this line of reasoning, since few of the changes in textual affiliation identified by them in fact involve any great stemmatic distance, which suggests relative constancy also of that variable. Phylogenetic analyses of the textual variations similarly show that the relationships between the manuscripts varies little between the various parts of the poem (Robinson 1997, 2000, 2004, 2006; Robinson and Bordalejo 2006). To decide whether the conversion of spelling forms was partial or complete, it was of interest to establish how N-gram models would fare with a more diverse corpus.

Multiple scribes, multiple authors

The manuscript preserved as Edinburgh, National Library of Scotland, Advocates' 19.2.1 ('Auchinleck') comprises forty-four texts, of which all but one are versified. It is executed in six scribal hands such that a change of hand always falls at a textual boundary (Bliss 1951; Cunningham 1972; Pearsall and Cunningham 1977; Wiggins 2004). It is considered a London product, and the spelling of Scribes 1 and 3 shows sufficient similarity to have been viewed as an example of beginning standardisation in the capital (Samuels 1963; cf. Runde 2010). Scribe 6's paleographical profile resembles that of Scribe 1 but his linguistic profile differs in numerous characteristics and is localisable to the south-west Midlands (Wiggins 2004), near that of Scribe 2. Dialectologists place Scribe 5's profile in Essex.[4] Finally, Scribe 4's stint is the one non-versified text: it is a list of Norman surnames and as such not written in English at all. Several of the texts survive uniquely or in their earliest

[4] For the respective linguistic profiles for Scribes 1, 2, 3, 5, and 6 and their localisations, see McIntosh et al. (1986): LPs 6510, 6940, 6500, 6350, and 7820.

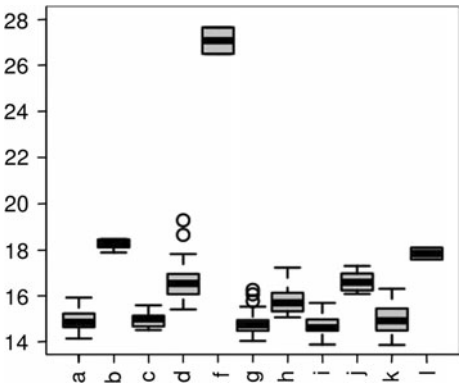

Fig. 1 Perplexity distribution in the Auchinleck manuscript. The horizontal axis gives group of segments in manuscript order. The vertical axis gives perplexity

known English-language copy in this manuscript but they are accepted as being products of several authors and translators. Various variables thus link segments of the Auchinleck text to each other. Similarity of spelling could align with any of these variables, whether singly or in combination.

The methodology for the second study was as follows. A transcript of the Auchinleck manuscript obtained from the Oxford Text Archive was segmented at every 200th line,[5] and each model trained on a separate segment by means of the SRILM toolkit. As in the first study, all punctuation was removed and all text was 'lowercased'. A spelling form was defined as a sequence of one or more letters occurring between two spaces. The training and test data both included grams spanning the space separating consecutive forms but excluded grams containing a line-boundary marker. Any final segment containing fewer than 200 lines was excluded. The models were again Witten-Bell smoothed, interpolated 3-gram models, and every model was tested on every segment. The resulting perplexity distribution had a positive skew of 10.2630, necessitating its log-transformation.[6] For every model, the mean of its log-transformed perplexity and associated standard deviation were established and visualised by means of a scatterplot. Groups of data points were next identified visually in the scatterplot. This process was repeated with several other segment sizes and with odd and even lines modelled separately to ensure that the groups were no artefact of the method or property of the texts' lexicon. The proposed groups were then further isolated through exclusion of every 200-line segment falling at a transition between two groups. Lastly, they were subjected to a one-way ANOVA test in conjunction with Tukey's Range Test by means of the R software environment for statistical computing.

Figure 1 above shows a boxplot of the perplexities computed for 200-line segments taken from the Auchinleck manuscript and arranged into groups. The vertical axis gives perplexity, while the horizontal axis gives the groups in their manuscript order. A box ends at the twenty-fifth and seventy-fifth quartiles with the horizontal line inside a box marking the statistical median. A T-bar stretches 1.5

[5] http://ota.ahds.ac.uk/headers/2493.xml. Accessed 25 December 2011. The transcript is also available in Burnley and Wiggins (2003) and includes silent expansion of abbreviated forms.

[6] Log-transformation reduced the skew to 5.1135.

times the interquartile range so that an upright T-bar, a box, and a downward T-bar together have a length of four times this range. Any higher or lower value is considered as an outlier and is represented by a circle.

Table 1 below gives, separately for each group, the probability that its mean perplexity is identical to that of another group by coincidence according to Tukey's Range Test. The bracketed numbers in the leftmost column identify the scribe of all text in a group, as the groups of 200-line segments were found to coincide with the scribal stints. The omission of segments falling at a transition left scribe 2's second stint altogether out of consideration. This stint is exclusively made up of the 98-line poem The Sayings of the Four Philosophers, which additionally contains a handful of lines in French; it is included as group e in Table 1 to indicate its place in the manuscript. It can be seen from the table that the statistics largely support the groups.

The qualification 'largely' is appropriate, since the models discriminate suboptimally. Many of the probabilities are \.0001, for example those separating Scribes 1 and 3; this is a value indicating a very clear difference. However, while Scribe 1's stints do cluster together, their pairwise comparisons could be expected to have resulted in probabilities closer to 1.0000 than they did; values in this region would have indicated a minuscule likelihood of the stints being similar by coincidence. The respective stints of Scribes 3 and 6 are similar by this value (P = 1.0000), and Scribe 6's stint j does not significantly differ from Scribe 2's stint l.

However, ability to discriminate is of course relative to the particular variable being tested. What is perceived as comparatively poor discrimination of this variable is simply due to interference from other variables. One reason that the experimental design was suboptimum for distinguishing the Auchinleck scribes was the exclusion of a line-boundary marker. Its exclusion may have blurred an important distinction, since scribes are known more frequently to reproduce the

Table 1 Pairwise comparisons among mean perplexities for groups of segments in the Auchinleck manuscript

	a	b	c	d	e	f	g	h	i	j	k
a (1)											
b (2)	\.0001										
c (1)	1.0000	\.0001									
d (3)	\.0001	\.0001	\.0001								
e (2)	–	–	–	–							
f (4)	\.0001	\.0001	\.0001	\.0001	–						
g (1)	.8962	\.0001	.9303	\.0001	–	\.0001					
h (5)	\.0001	\.0001	\.0001	\.0001	–	\.0001	\.0001				
i (1)	.6748	\.0001	.7907	\.0001	–	\.0001	1.0000	\.0001			
j (6)	\.0001	\.0001	\.0001	1.0000	–	\.0001	\.0001	.0027	\.0001		
k (1)	.9931	\.0001	1.0000	\.0001	–	\.0001	.2340	\.0001	.0722	\.0001	
l (2)	\.0001	.9964	\.0001	.0403	–	\.0001	\.0001	\.0001	\.0001	.0806	\.0001

spelling forms of the exemplar in line-initial and line-final positions in verse.[7] Of greater consequence was the inclusion of grams spanning the space between consecutive spelling forms. Their inclusion shifts some of the focus away from the spelling forms and toward the lexicon. The effect is anything but negligible since as many as about one in every six occurrences of a 3-gram spans a form boundary in a regular discursive text. The considerably higher perplexities reported in Fig. 1 above as compared with Fig. 2 below are due to these two methodological issues.

One scribe, one author

Still, it is unambiguous that the perplexities do principally divide up the text of the Auchinleck manuscript by scribal stint. This division again suggests conversion of spelling forms so thorough as to have erased every trace of the exemplars. It would, however, be hasty so to generalise about scribes. Partial conversion leads to the same groups, provided more spelling forms have been introduced by the individual scribe than have been retained from the exemplars, and a partial conversion is what the Auchinleck data show. This conclusion can best be appreciated by studying a longer text executed in a single scribal hand. The final preparatory study was of such a text but was likewise conducted at a comparatively early stage in the development of the methodology of quantifying the similarity of spelling between medieval English texts by means of N-gram models (Thaisen, in press).

The training and test data for this study were the copy of the Canterbury Tales found in Cambridge University Library, MS Gg.4.27, which dates from late in the first quarter of the fifteenth century. The study was construed as a pilot of the full-scale study discussed below and partially drew on the same corpus. A transcript of Gg.4.27 was pruned of punctuation, 'lowercased', and segmented. Each segment comprised 40,000 characters in a first series so as to keep the size constant. This division gave a total of no more than nineteen segments, while a second series divided up the poem by tale, achieving twenty-five segments of varying size. Every segment was separately modelled using SRILM. The models were smoothed and interpolated in the same manner as in the previous studies. For the reasons discussed above, they again included grams spanning the space separating consecutive forms but excluded grams containing a line-boundary marker. Every model was tested on every segment in the two separate series. Mean perplexity and associated standard deviation were established for every model and plotted on a scatterplot. Visual inspection of the scatterplot suggested three groups, but statistical testing confirmed the existence of only two groups.

There was outside support for the boundaries of the groups coinciding with the boundaries of possible exemplars, as follows. The two populations met at the junction of the Summoner's and Clerk's Tales; this junction marks the end of what the Riverside edition of the poem calls Fragment D (Benson 1987). The group whose independence could not be statistically verified consisted of the tales up to and including the Reeve's Tale; this tale concludes what the Riverside edition calls Fragment A, since the Cook's Tale is lost from Gg.4.27 through mutilation. Changes

[7] Restricting the enquiry exclusively to line-initial and line-final positions would have prohibitively reduced the number of grams collected from a sample.

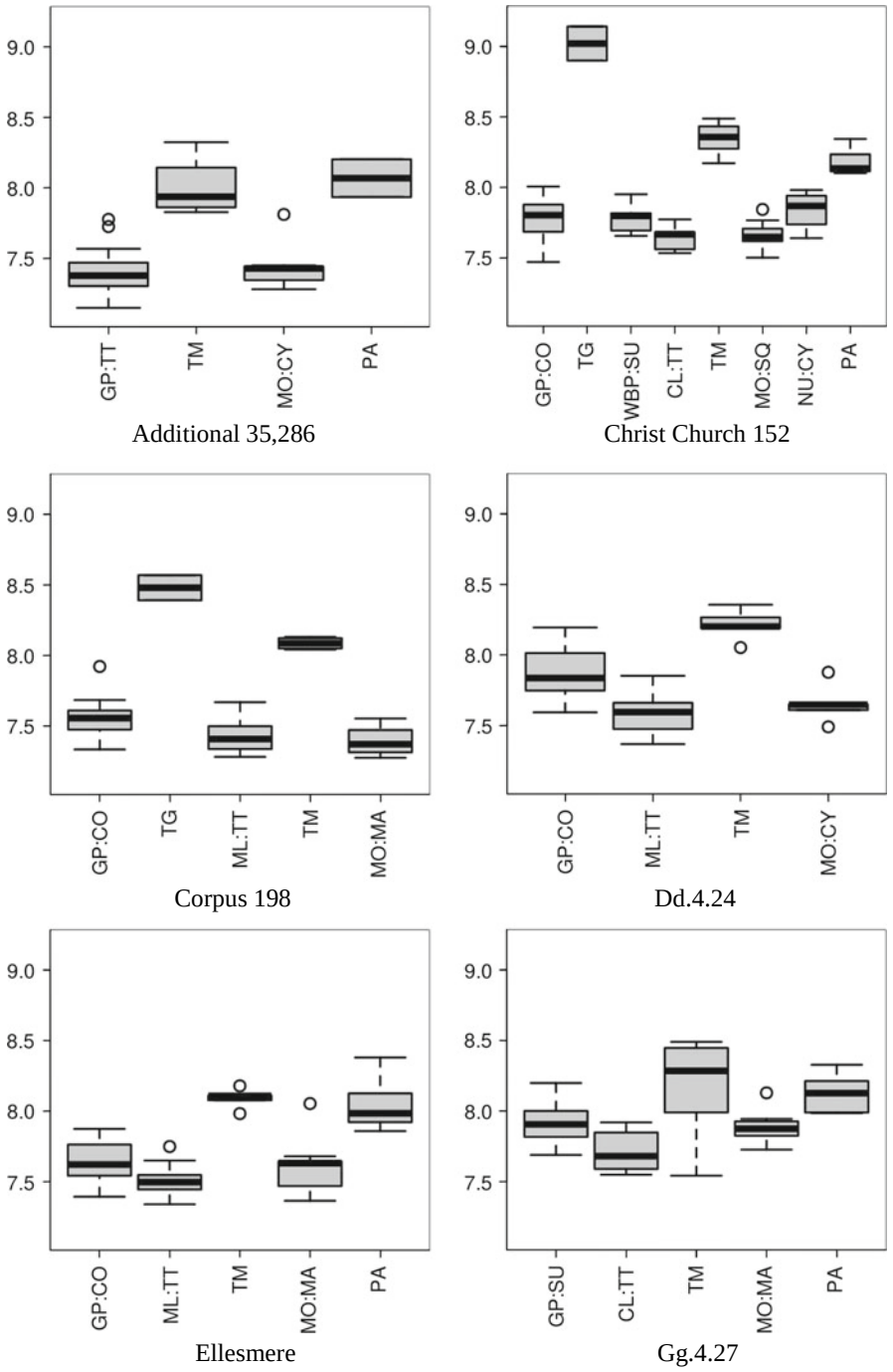

Fig. 2 Perplexity distribution in nine manuscripts of Chaucer's Canterbury Tales. The horizontal axis gives group of segments in manuscript order. The vertical axis gives perplexity

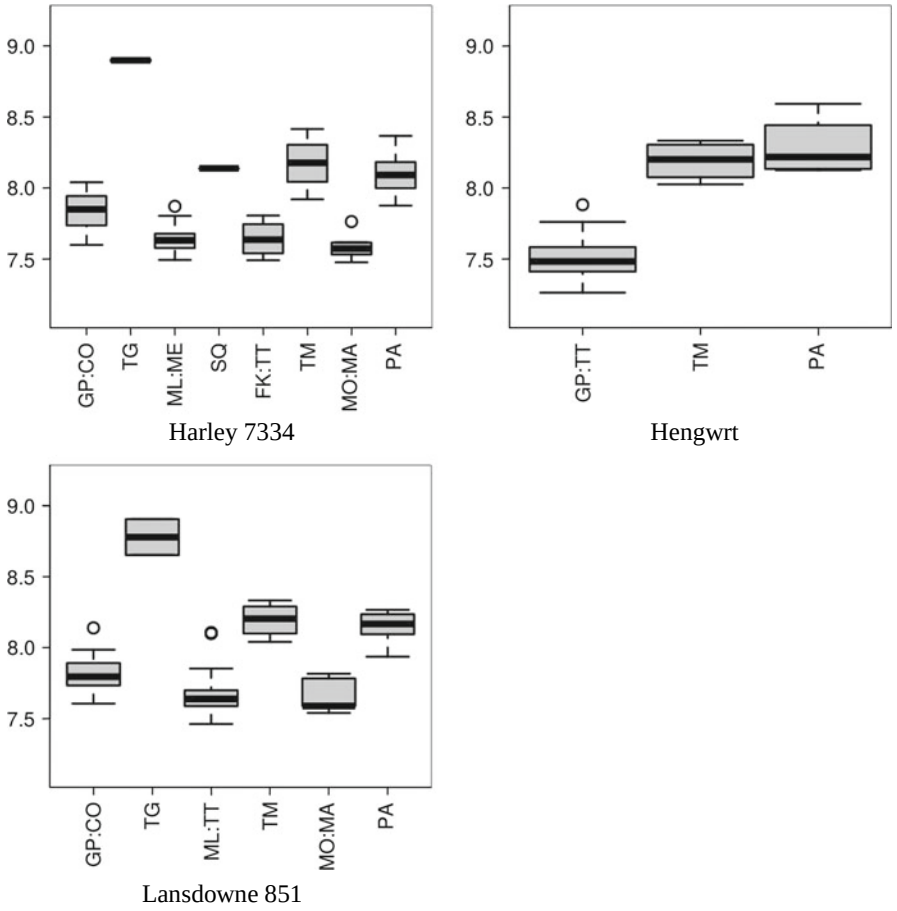

Fig. 2 continued

in ink, ruling, and quire signatures converged on these two locations, and the sole change in textual affiliation involving any noteworthy stemmatic distance fell at the end of Fragment D. The discussion concluded on this basis that the number of groups corresponded to the number of scribal hands responsible for the exemplars. This conclusion, then, implied what the comparative study of all copies of the Miller's Tale and the Wife of Bath's Prologue also implied: that if the text of the Canterbury Tales reached a scribe as a series of physically distinct exemplars, these exemplars may have gone back to a small number of scribes and therefore probably never circulated independently of each other.

The Canterbury Tales manuscripts with early Contents

Previous studies have, in other words, validated perplexity of smoothed, interpolated N-gram models of spelling as a similarity metric. Similarity of spelling between texts, in turn, indicates relative constancy of the variables capable of skewing a scribe's spelling in one direction or another. The scribe's own spelling practices are the strongest of these variables, followed by the practices of the respective scribes of the exemplars. Each of the manuscripts of

the Canterbury Tales with early textual contents was copied in its entirety by a different scribe, although two of the scribes each copied two of the manuscripts. A study of these manuscripts may accordingly reveal the number of scribes behind their exemplars and in turn indicate the level of integration of the Tale of Gamelyn. To ascertain this number, I obtained transcripts of nine manuscripts of the Canterbury Tales courtesy of the Canterbury Tales Project. As before, the transcripts adhered to the transcription conventions devised by Robinson and Solopova (1993), which are standard for the Project and which involve little editorialising. The manuscripts were all those considered stemmatically central because of their early contents: Christ Church 152, Corpus 198, Ellesmere, Gg.4.27, and Hengwrt have already been mentioned. Add to these Cambridge, University Library, MS Dd.4.24 and London, British Library, MSS Additional 35,286 and Lansdowne 851.[8]

My methodology was a refinement of that already described. Extensive testing, including the three studies reported above, led me to conclude that changes in spelling in a manuscript are best captured with each transcript partitioned into 2,700-form segments pruned of punctuation, all text 'lowercased', all grams spanning the space separating consecutive spelling forms excluded, and a line-boundary marker included; this segment size corresponds to about 10,000 3-grams. The specific smoothing technique is immaterial with data volumes this sparse (Chen and Goodman 1998), and gram lengths of 2 and 3 discriminate equivalently. The transcripts segmented accordingly, each segment served as training data for a separate Witten-Bell smoothed, interpolated 3-gram model. For each manuscript separately, every model was tested on every segment from the same manuscript, resulting in a series of perplexities. For every model, its mean perplexity and associated standard deviation were established along with the skew of the perplexity distributions; none of the distributions had any significant negative or positive skew, so no step was taken to normalise them further.[9] The perplexities were once again visualised by means of a scatterplot and visually inspected.

The inspection suggested the groups of segments apparent from Fig. 2 above. The groups have been further isolated through exclusion of every segment coinciding with a transition between two groups. This policy explains the absence of three short items: the Ploughman's Tale from the groups identified in Christ Church 152, the Parson's Tale from those identified in Corpus 198, and the Parson's Prologue from those identified in Hengwrt, although all three coincide with discontinuities in the scatterplots. The figure gives boxplots of perplexities based on

[8] There are two exceptions to the statement that each of these manuscripts was copied in its entirety by a different scribe. The Ploughman's Tale is an addition to Christ Church 152 executed in a later hand, and a different hand is responsible for fols. 508^{r}–510^{v} and 514 recto and verso in the final quire of Gg.4.27, which contains John Lydgate's Temple of Glas.

[9] The respective skews of the perplexity distributions were -.2363 for Additional 35,286, .4702 for Christ Church 152, -.3283 for Corpus 198, -.1790 for Dd.4.24, -.4771 for Ellesmere; -.2243 for Gg.4.27, -.0820 for Harley 7334, -.0973 for Hengwrt, and -.0836 for Lansdowne 851.

2,700-form segments taken from nine manuscripts of Chaucer's Canterbury Tales and arranged into groups. The plots have the same layout as Fig. 1 above. Poetic form had appeared from the pilot study of Gg.4.27 to be a non-salient variable, but application on the full corpus revealed that the manuscript is untypical in this respect. A contrast between verse and prose is in evidence, as will become clear. However, removal of prose from the corpus and recalculation of perplexity produced virtually no change in the relationships between the verse parts.[10] Because of this inconsequentiality, both poetic forms are included in what follows so as to present an analysis of the complete text of every manuscript. Lastly, the R software was used to subject the groups to a one-way ANOVA test in conjunction with Tukey's Range Test.

Table 2 below gives the probability that the groups' mean perplexities could coincidentally be identical in the context of the individual manuscript according to Tukey's Range Test.

There is recurrence across the manuscripts in what groups are distinguished by the statistical tests presented in Table 2. As a first group, the tests show that the Tale of Gamelyn ('TG') always is significantly different from all other text ($P \leq .01$). The tale is present in Christ Church 152, Corpus 198, Harley 7334, and Lansdowne 851, but absent from the remaining five manuscripts.

As a second group, the tales of Melibee ('TM') and the Parson ('PA') do not differ significantly from each other but do both differ significantly from all remaining text in the individual manuscript ($P \leq .01$). This is the case with Additional 35,286, Christ Church 152, Ellesmere, Hengwrt, and Lansdowne 851. The Tale of Melibee differs significantly from all remaining text in Corpus 198 and Dd.4.24. A fragmentary Parson's Tale occupies fols. 262^{r}–266^{v} in the former of these two manuscripts but is excluded from the data due to the principle of excluding segments either containing less than 2,700 forms or falling at a transition, as has already been mentioned. The tale is absent from the latter manuscript due to loss of leaves. Finally, the tales of Melibee and the Parson both differ significantly from all remaining text in Harley 7334 except the Squire's Tale ('SQ'), which in turn does not differ significantly from Fragment A ('GP:CO').

Thirdly, there is support for the independence of a group of canonical verse tales from other such groups within a single manuscript. Disregarding the Tale of Gamelyn, Fragment A ('GP:CO') differs significantly from the group of canonical tales following it in Corpus 198, Dd.4.24, and Harley 7334 ($P \leq .01$). The fragment is marginally significantly different from the tales following it in Ellesmere ($P \leq .05$). Lastly, Fragment A, the Man of Law's Tale, and Fragment D together ('GP:SU') marginally significantly differ from a group of tales stretching from the Clerk's Tale to the Tale of Sir Thopas ('CL:TT') in Gg.4.27 ($P \leq .05$).

Visual analysis had suggested no more than two further sudden changes in perplexity in any of the manuscripts. Neither is significant. Not significantly different is thus Fragment A ('GP:CO') in Lansdowne 851 from the consecutive group of canonical verse tales following it, which extends from the Man of Law's

[10] Removal of prose altered the perplexities for the verse parts by no more than .0228 on average with a standard deviation of .0060.

Table 2 Pairwise comparisons among mean perplexities for groups of segments in nine manuscripts of the Canterbury Tales

	GP:TT	TM	MO:CY
Additional 35,286			
TM	\.0001		
MO:CY	1.0000	\.0001	
PA	\.0001	.6579	\.0001

	GP:CO	TG	WBP:SU	CL:TT	TM	MO:SQ	NU:CY
Christ Church 152							
TG	\.0001						
WBP:SU	1.0000	\.0001					
CL:TT	.0681	\.0001	.2916				
TM	\.0001	\.0001	\.0001	\.0001			
MO:SQ	.0601	\.0001	.3423	.9999	\.0001		
NU:CY	.9920	\.0001	.9819	.0853	\.0001	.1006	
PA	\.0001	\.0001	\.0001	\.0001	.3229	\.0001	.0005

	GP:CO	TG	ML:TT	TM
Corpus 198				
TG	\.0001			
ML:TT	.0046	\.0001		
TM	\.0001	.0015	\.0001	
MO:MA	.0703	\.0001	.9849	\.0001

	GP:CO	ML:TT	TM
Dd.4.24			
ML:TT	\.0001		
TM	.0006	\.0001	
MO:CY	.0761	.7158	\.0001

	GP:CO	ML:TT	TM	MO:MA
Ellesmere				
ML:TT	.0252			
TM	\.0001	\.0001		
MO:MA	.9522	.2994	\.0001	
PA	\.0001	\.0001	.9462	\.0001

	GP:SU	CL:TT	TM	MO:MA
Gg.4.27				
CL:TT	.0158			
TM	.0945	.9949		
MO:MA	.1124	.0003	.2242	
PA	.0002	.1127	.9996	.1376

Table 2 continued

	GP:CO	TG	ML:ME	SQ	FK:TT	TM	MO:MA
Harley 7334							
TG	\.0001						
ML:ME	.0010	\.0001					
SQ	.3691	.0003	.0083				
FK:TT	.1424	\.0001	1.0000	.0234			
TM	.0003	\.0001	\.0001	1.0000	\.0001		
MO:MA	.0122	\.0001	.9964	.0068	.9989	\.0001	
PA	.0015	\.0001	\.0001	1.0000	\.0001	.9444	\.0001

	GP:TT	TM
Hengwrt		
TM	\.0001	
PA	\.0001	.5275

	GP:CO	TG	ML:TT	TM	MO:MA
Lansdowne 851					
TG	\.0001				
ML:TT	.0828	\.0001			
TM	.0006	.0004	\.0001		
MO:MA	.4021	\.0001	1.0000	\.0001	
PA	\.0001	\.0001	\.0001	.9968	\.0001

Tale up to and including the Tale of Melibee (‘ML:TM’). Nor significantly different is Fragment D (‘WBP:SU’) from the consecutive group of tales from the Clerk’s Tale up to and including the Tale of Sir Thopas (‘CL:TT’) in Christ Church 152. Fragment A (‘GP:CO’) could, however, be considered marginally significantly different from this group (P = .0681).

These similarity metrics would appear to suggest the possibility that a greatly restricted number of scribes produced the respective exemplars for these nine manuscripts of the Canterbury Tales with early textual contents. This number seems on this evidence to have been no more than one for Additional 35,286, Christ Church 152, Hengwrt, and Lansdowne 851, one or two for Ellesmere, and two for Corpus 198, Dd.4.24, Gg.4.27, and Harley 7334, as far as the canonical verse tales are concerned. The results also show the unforeseen variable of poetic form to be at play, as the separateness of the tales of Melibee and the Parson may reasonably be attributed to their constituting the sole portions in prose, rather than to any difference of scribe at the level of the exemplars. As for the noncanonical Tale of Gamelyn, for each manuscript, a separate scribe may be responsible for its exemplars: there are reasons, therefore, for putting forward the possibility that this tale might be the only tale with a history of its own.

Discussion

It is the earliest and most authoritative surviving copies of the Canterbury Tales which some scholars now associate with scribes employed at the London Guildhall on paleographical evidence. Among them, the Tale of Gamelyn is present in Corpus 198 and Harley 7334, which are products of one scribe. It is absent from two further manuscripts, Hengwrt and Ellesmere, which are products of another scribe. This second scribe may have made at least one further copy of the Canterbury Tales, as a three-leaf fragment of the Nun's Priest's Tale in his hand survives as Aberystwyth, National Library of Wales, 21972D.[11] Their respective names may have been John Marchaunt and Adam Pinkhurst (Mooney 2006; Mooney & Stubbs, in press; cf. Roberts 2011). They were formerly known as, respectively, Scribes D and B, these labels reflecting the order of their appearance in Cambridge, Trinity College, MS R.3.2, which houses a copy of John Gower's Confessio Amantis to which they both contributed (Doyle and Parkes 1978). Their appearance there has long suggested to scholars that they were acquainted and possibly regularly collaborated. What directly associates these two scribes with the Guildhall are their contributions to Letter Books H and I preserved in Kew, The National Archives, COL/AD/01 (Mooney and Stubbs, in press). This association further supports the suggestion that they knew one another, as does such less direct evidence as their copies of the Canterbury Tales being linked to each other through their decoration (Rickert 1940; Scott 1995).

Despite the proximity of Scribes B and D to each other, their four complete manuscripts of the Canterbury Tales differ in the inclusion and ordering of canonical contents. The locations of text at which these differences are manifest often coincide with boundaries in the physical makeup of the individual manuscript, implying that the manuscripts were non-consecutively copied from exemplars only partially ordered. It is easy to picture how a scribe in the process of making sense of exemplars in this format could have incorporated the Tale of Gamelyn after the Cook's Tale and another have had no access to the Canon's Yeoman's Tale. An example is visible in the perplexity distribution for Harley 7334. Its Squire's Tale originally followed the Man of Law's Tale but was allocated to a position after the Merchant's Tale after it had been copied (Blake and Thaisen 2004).[12]

It may have been the norm in medieval London for longer texts to be parcelled out to multiple scribes for simultaneous copying. An example is Trinity R.3.2, and coincident textual and codicological boundaries suggest the manuscripts of Chaucer's Troilus and Criseyde as another example (Hanna 1996). It has been

[11] Doyle and Parkes (1978), Doyle (1995), and Mosser (1996) discuss whether Cambridge, University Library, Kk.1.3, part 20, a fragment of the Prioress's Tale, also is written in this scribe's hand.

[12] See catalogues such as Manly and Rickert (1940) and Mosser (1996) for further examples. Stubbs (2007) describes a previously unnoticed replacement of folios in Corpus 198. Particularly important for Ellesmere is the division of its quires by border artist presented in Scott (1995). With the Squire's Tale in its original position Harley 7334 would order the tales identically to Corpus 198, the original position being after the Man of Law's Tale, which is a feature also found in Hengwrt. Disregarding links, the move brought the order closer to that found in Ellesmere, the differences being in the inclusion of the Tale of Gamelyn and, further on in the poem, the placement of Fragment G, which comprises the tales of the (Second) Nun and Canon's Yeoman.

proposed that the text of the Canterbury Tales too was so transmitted (Horobin 2010), the partially ordered format of the exemplars giving rise to the differences in tale order in evidence between the various manuscripts with early textual contents.

However, this proposed mode of transmission may not entirely accord with the integrity of the exemplars suggested by the similarity metrics. It is not pushing the bounds of credulity to suggest that no other exemplars may have been available than those relied on by Scribes B and D in producing their manuscripts. Only two complete copies of the poem appear to survive from the next decade: Dd.4.24 and Lansdowne 851. The former manuscript contains the Nun's Priest's Endlink (lines $B^2$4637–52) usually considered canonical but absent from every manuscript associated with Scribes B and D; its presence is what differentiates the a order from the Ellesmere order. Dd.4.24 is otherwise close to Ellesmere in both text and spelling (Manly and Rickert 1940, i:102; Robinson 1997, 2000, 2004; Horobin 2003:149; Thaisen and Da Rold 2009). The perplexity distribution reveals two scribes to be behind its exemplars, the boundaries of their stints coinciding with those in Ellesmere. No irrefutable evidence has yet been unveiled to link Wytton, Dd.4.24's scribe, firmly to Scribes B and D but his paleography and his manuscript's codicological features–for example, the date of its watermarks–probably do situate its production in a similar environment in the London area around the turn of the fifteenth century (Da Rold 2003, 2007; Thaisen and Da Rold 2009).

The link is stronger for Lansdowne 851. This manuscript has a close textual relationship with Corpus 198, probably by way of a shared ancestor (Manly and Rickert 1940, i:306; Blake 1985: 73; Robinson 1997, 2000, 2004; Thomson 1998: 235–236), and may have been decorated by the artisans responsible for decorating Corpus 198, Ellesmere, and Harley 7334 (Scott 1995: 104 and 117, n.44; cf. Rickert 1940: 568). Manly and Rickert, who held Corpus 198 and Lansdowne 851 to be products of 'the same scribal and illuminating shop', noted possible kinship relations between their early owners (1940,i:98–99 and 307). The scribal hand of Lansdowne 851 has the single-compartment a, right-shouldered r, and kidney-shaped s characteristic of Secretary script, which may possibly further associate the manuscript with the type of environment in which Scribes B, D, and Wytton were active (Mosser 1996; Seymour 1997: 134; Thomson 1998: 59–62).[13] It would appear, therefore, that neither of these two manuscripts can be confidently dissociated from Scribes B and D.

Also possibly important may be that the scribes of exemplars discernible in the similarity metrics describing Lansdowne 851 are statistically only marginally significantly different from each other. The reason for the possible importance is that these blurred boundaries are consistent with one or more preceding layers of scribal copying, each with partial conversion of the spelling forms found in the exemplars. The boundaries of Fragment D are equally blurred in the similarity metrics respectively describing Christ Church 152 and Gg.4.27, which likewise are not stemmatically primary. Previous scholarship holds the final manuscript with

[13] See Horobin (2003: 67–68, cf. 152–153) for the view that Lansdowne 851 was produced in the south-west Midlands.

early textual contents, Additional 35,286, to have been consecutively copied from a single exemplar (Manly and Rickert 1940, i:43; Horobin 1997, i:64–66, 235–236; Thaisen 2008b), and the similarity metrics make it probable that this exemplar was written in a single scribal hand. It is known that few of the later complete manuscripts are written in more than a single hand,[14] and it may be recalled how phylogenetic analyses of the textual variations essentially suggest a stable set of relationships between all the manuscripts throughout the text. There are possible grounds, therefore, for imagining that the text of the poem normally became transmitted as a single whole almost as soon as it had left the hands of Scribes B and D, rather than parcelled out for simultaneous copying.

Returning to the Tale of Gamelyn, the separate scribe for its exemplars would disfavour any possible Chaucerian authorship. The harder question would appear to be why the tale is present in only two of Scribes B and D's manuscripts, specifically Scribe D's two manuscripts. It may be recalled that recent scholarship has, in view of the stemma, read their manuscripts' codicology as bearing witness to Chaucer adding to, revising, and rearranging the Canterbury Tales. However, the well-known annotation 'Of this Cokes tale maked Chaucer na moore' on Hengwrt's fol. 57^{v} suggests that the poet passed away before this manuscript's completion,[15] and the likely posthumousness of this annotation is one of several reasons why other scholars have seen a greater distance to him. Their studies have typically recognised that Scribe B had access to authoritative exemplars, while remaining silent on whether Scribe D did too.[16]

It is possible that the presence of two scribes is a feature only of the exemplars for Scribe D's manuscripts among the four. This is because the difference in perplexity is no more than marginally statistically significant between Fragment A and the verse tales following it in Ellesmere. This marginality suggests an intervening layer of copying but a more parsimonious reading is possible in view of Ellesmere's early production date and its position near the root of the stemma, unlike in the cases of Christ Church 152, Gg.4.27, and Lansdowne 851: that this layer conceivably records Scribe B copying from exemplars prepared by himself to a greater extent than is true of any other portion of Hengwrt and Ellesmere.

If this attribution is accepted, it could possibly rather be that Scribe B passed exemplars to Scribe D for him to copy from in his own workshop, all the while that they may have discussed the overall ordering. Any closer collaboration between Scribes B and D could be expected to have led to their hands alternating more frequently in the exemplars than they do. If what is found in these manuscripts does result from collaborative effort with or without Chaucer standing on the side, it would appear that either Scribe B omitted the Tale of Gamelyn or Scribe D inserted it. Blake (2004) argues the former: Scribe B omitted the Tale of Gamelyn from

[14] The most notable exception is London, British Library, MS Harley 7335, which is executed in no less than five hands. This is a manuscript with probable roots in Leicester and dating from the third quarter of the fifteenth century.

[15] A similar annotation 'Here endith the Squyeres tale as meche as Chaucer made' appears on Dd.4.24's fol. 126^{r} preceding a gap.

[16] Mosser (2008) summarises both the evidence and the debate. Other recent contributions are Stubbs (2007), Horobin (2010), Mooney and Stubbs, in press.

Ellesmere to prevent the occurrence of a singleton quire to take the final fourteen lines of the tale. Blank, but ruled folios corresponding to a regular quire of eight today constitute Ellesmere's flyleaves, and the manuscript's sole longer gap occurs after the Cook's Tale where it stretches to the end of the quire. It is the space provided by this gap and what is now the flyleaves which together are insufficient for taking the tale.

However, it may be possible to see something other than a local motivation for the absence of the Tale of Gamelyn from Ellesmere. The tale is present in the three manuscripts with the c order, defined by Corpus 198, as well as in the numerous ones with the d order, which is a later development of the c order. No later manuscript with the a-Ellesmere order or its derivative, the b order,[17] seems to derive from Ellesmere itself according to textual scholars but rather from closely related materials (Manly and Rickert 1940; Robinson 1997, 2000, 2004, 2006; Robinson and Bordalejo 2006); yet all other manuscripts with the a or b order too have no Tale of Gamelyn, including the near-contemporary Dd.4.24 where the presentation of the text was less of a concern.[18]

There are two exceptions. The tale occurs in the much-later Christ Church 152, whose tale order is unique but related to the a order. There it occupies an original blank after the Cook's Tale in quire 3 plus the irregular, added quire 4. The final manuscript to house the Tale of Gamelyn is Harley 7334. In it, the tale similarly occupies an original blank after the Cook's Tale plus an irregular quire finishing in another blank and was in the assessment of Manly and Rickert (1940, i:95–96; cf. ii:41–44) 'visibly picked up from Cp [Corpus 198]'. This distribution would suggest that the tale may not have been present in the exemplars behind the a-Ellesmere manuscripts, which in turn may probably indicate that Scribe D inserted it. What the manuscripts would seem to show, therefore, is Scribes B and D responding differently to exemplars not fully ordered, as opposed to Chaucer changing his mind.

Consulting with Scribe B about how to order these materials, Scribe D may have received an ordered Fragment A from him along with the Squire's Tale which went into Harley 7334. Like later manuscripts with the a-Ellesmere order relative to Ellesmere, the later c manuscripts appear not to have Corpus 198 as their shared ancestor, nor do the later d manuscripts (Manly and Rickert 1940; Robinson 1997, 2000, 2004, 2006; Robinson and Bordalejo 2006), and the similarity metrics have indicated how the text of the poem possibly soon came normally to be transmitted as a whole. Perhaps Scribes B and D each developed his own set of exemplars. Perhaps these sets became the respective ancestors of the a-Ellesmere and c traditions. Perhaps these ancestors are, in turn, identical to the a and c hyparchetypes recently

[17] The manuscripts defining the b and d orders are respectively Princeton, Firestone Library 100 ('Helmingham') and Petworth, The National Trust, Petworth House 7. Their orders not being primary is the reason why the two manuscripts were not included in the present study.

[18] Presentation must have been less of a concern, since the scribe left five gaps in Dd.4.24. He presented the text in a single column with a variable 41–48 lines to the page written within a frame. No ruling for lines is found, and cramping is evident toward the bottom of a number of pages. The manuscript is not illuminated.

posited on the basis of phylogenetic analyses of the textual variations (Robinson 2004, 2006).

Conclusion

Complete scribal conversion of exemplars' spelling forms is improbable. Try as a scribe might to effect a complete conversion, he will always have been subconsciously primed by the spelling forms he found there. The result was a skew of his unconstrained selection of spelling forms in the direction of the exemplars. Other scribes will have set out to produce a partial conversion, leading to a similar result. Still other scribes will have aimed at reproducing their exemplars literatim, in which case their copy can be treated as if it was its exemplars. This skew provides a window on the number of scribes behind the exemplars for a Middle English text executed in a single hand. Standard techniques from natural language processing reduce another skewing effect, that of a sample's lexicon on its spelling profile, to permit objective quantification of the similarity between spelling profiles such as N-gram models. Their comparison reveals the number of previous scribes, provided allowance is made for variations in spelling that are due to poetic form. This number would seem never to have been more than three for any individual manuscript of Chaucer's Canterbury Tales with an early text.

Acknowledgments The author thanks Nina Haver, Peter Robinson, Merja Stenroos, and Estelle Stubbs for helpful comments on an earlier draft, and Brian Roark and Richard Sproat for advice on probabilistic modelling. The author was the grateful beneficiary of a stipend from the Medieval Institute at the University of Notre Dame, which facilitated the preparation of the paper.

References

Benskin, M. & Laing, M. (1981). Translations and Mischsprachen in Middle English manuscripts. In M. Benskin & M. Samuels (Eds.), So meny people longages and tonges: Philological essays in Scots and mediaeval English presented to Angus McIntosh (pp. 55–106). Edinburgh: Middle English Dialect Project.

Benson, L. (Ed.) (1987). The Riverside Chaucer (3rd ed.). Boston: Houghton Mifflin.

Blake, N. (1985). The textual tradition of the 'Canterbury Tales'. London: E. Arnold.

Blake, N. (1997). Geoffrey Chaucer and the manuscripts of the Canterbury Tales. Journal of the Early Book Society, 1, 96–122.

Blake, N. (2004). Chaucer, Gamelyn, and the Cook's Tale. In T. Matsuda, R. Linenthal, & J. Scahill (Eds.), The medieval book and a modern collector: Essays in honour of Toshiyuki Takamiya (pp. 87–98). Cambridge: D. S. Brewer.

Blake, N. & Thaisen, J. (2004) Spelling's significance for textual studies. In C. Dollerup (Ed.), Worlds of words: A tribute to Arne Zettersten (pp. 93–107) [Nordic Journal of English Studies, special issue, 3(1)].

Bliss, A. (1951). Notes on the Auchinleck manuscript. Speculum, 26, 652–658.

Burnley, D. & Wiggins, A. (Eds.) (2003). The Auchinleck manuscript, version 1.1. The National Library of Scotland. http://auchinleck.nls.uk/. Accessed 24 December 2011.

Chen, S. & Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. Technical report TR-10–98, Harvard University. http://research.microsoft.com/joshuago/tr-10-98.pdf. Accessed 24 December 2011.

Cunningham, I. (1972). Notes on the Auchinleck manuscript, Speculum, 47, 96–98.

Da Rold, O. (2003). The quiring system in Cambridge University Library MS. Dd.4.24 of Chaucer's Canterbury Tales. The Library, 4, 107–128.
Da Rold, O. (2007). The significance of scribal corrections in Cambridge University Library MS. Dd.4.24 of Chaucer's Canterbury Tales. The Chaucer Review, 41, 393–438.
Doyle, A. (1995). The copyist of the Ellesmere Canterbury Tales. In M. Stevens & D. Woodward (Eds.), The Ellesmere Chaucer: Essays in interpretation (pp. 49–67). San Marino, CA: Huntington Library.
Doyle, A. & Parkes, M. (1978). The production of copies of the Canterbury Tales and the Confessio Amantis in the early fifteenth century. In M. Parkes & A. Watson (Eds.), Mediaeval scribes, manuscripts and libraries: Essays presented to N. R. Ker (pp. 163–203). Aldershot: Scolar Press.
Hanna, R. (1989). The Hengwrt manuscript and the canon of the Canterbury Tales. English Manuscript Studies 1100–1700, 1, 64–84.
Hanna, R. (1996). Pursuing history: Middle English manuscripts and their texts. Stanford: Stanford University Press.
Horobin, S. (1997). A transcription and study of British Library MS Additional 35286 of Chaucer's Canterbury Tales. Unpublished Ph.D. dissertation, University of Sheffield.
Horobin, S. (2003). The language of the Chaucer tradition. Cambridge: D.S. Brewer.
Horobin, S. (2010). Adam Pinkhurst, Geoffrey Chaucer, and the Hengwrt manuscript of the Canterbury Tales. The Chaucer Review 44(4), 351–367.
Manly, J. & Rickert, E. (Eds.) (1940). The text of the 'Canterbury Tales': Studied on the basis of all known manuscripts [8 vols]. Chicago: University of Chicago Press.
McIntosh, A. (1963). A new approach to Middle English dialectology. English Studies, 44, 1–11.
McIntosh, A., Samuels, M., & Benskin, M. (Eds.) (1986). A linguistic atlas of late mediaeval English [4 vols]. Aberdeen: Aberdeen University Press.
Mooney, L. (2006). Chaucer's scribe. Speculum: A Journal of Medieval Studies, 81, 97–138.
Mooney, L. & Mosser, D. (2004). The hooked-g scribes and Takamiya manuscripts. In T. Matsuda, R. Linenthal, & J. Scahill (Eds.), The medieval book and a modern collector: Essays in honour of Toshiyuki Takamiya (pp. 179–196). Cambridge: D.S. Brewer.
Mooney, L. & Stubbs, E. Scribes and the city: London guildhall clerks and the dissemination of Middle English literature 1375–1425. Woodbridge: York Medieval Press (in press).
Mosser, D. (1996). Witness descriptions. In P. Robinson (Ed.), The Wife of Bath's Prologue on CD-ROM (n.p.). Cambridge: Cambridge University Press.
Mosser, D. (2008). 'Chaucer's Scribe', Adam, and the Hengwrt project. In M. Connolly & L. Mooney (Eds.), Design and distribution of late medieval manuscripts in England (pp. 11–40). York: York Medieval Press.
Owen, C. (1991). The manuscripts of the 'Canterbury Tales'. Cambridge: D. S. Brewer.
Pearsall, D. & Cunningham, I. (Eds.) (1977). The Auchinleck manuscript. London: Scolar Press.
Rickert, M. (1940). Illumination. In J. Manly & E. Rickert (Eds.), The text of the 'Canterbury Tales': Studied on the basis of all known manuscripts (vol. 1, pp. 561–605). Chicago: University of Chicago Press.
Roberts, J. (2011). On giving Scribe B a name and a clutch of London manuscripts from c.1400. Medium Aevum, LXXX, 231–254.
Robinson, P. (Ed.) (1996). The 'Wife of Bath's Prologue' on CD-ROM. Cambridge: Cambridge University Press.
Robinson, P. (1997). A stemmatic analysis of the fifteenth-century witnesses to the Wife of Bath's Prologue. In N. Blake & P. Robinson (Eds.), The Canterbury Tales Project Occasional Papers (pp. 69–132). Oxford: Office for Humanities Communication.
Robinson, P. (2000). Stemmatic commentary. In E. Solopova (Ed.), The 'General Prologue' on CD-ROM (n.p.). Cambridge: Cambridge University Press.
Robinson, P. (Ed.) (2004). The 'Miller's Tale' on CD-ROM. Leicester: Scholarly Digital Editions.
Robinson, P. (2006). Witness relations. In P. Thomas (Ed.), The 'Nun's Priest's Tale' on CD-ROM (n.p.). Birmingham: Scholarly Digital Editions.
Robinson, P. & Bordalejo, B. (2006). Stemmatic commentary. In P. Thomas (Ed.), The 'Nun's Priest's Tale' on CD-ROM (n.p.). Birmingham: Scholarly Digital Editions.
Robinson, P. & Solopova, E. (1993). Guidelines for transcription of the manuscripts of the Wife of Bath's Prologue. In N. Blake & P. Robinson (Eds.), The Canterbury Tales Project Occasional Papers (pp. 19–52). Oxford: Office for Humanities Communication.
Runde, E. (2010). Reexamining orthographic practice in the Auchinleck manuscript through study of complete scribal corpora. In R. Cloutier, A. Hamilton-Brehm, & W. Kretzschmar (Eds.), Studies in

the history of the English language V: Variation and change in English grammar and lexicon (pp. 265–287). Berlin: Walter de Gruyter.
Rybicki, J. & Eder, M. (2011). Deeper Delta across genres and languages: Do we really need the most frequent words? Literary and Linguistic Computing, 26(3), 315–321.
Samuels, M. (1963). Some applications of Middle English dialectology. English Studies, 44, 81–94.
Scott, K. (1995). An hours and psalter by two Ellesmere illuminators. In M. Stevens & D. Woodward (Eds.), The Ellesmere Chaucer: Essays in interpretation (pp. 87–119). San Marino, CA: Huntington Library.
Seymour, M. (1997). A catalogue of Chaucer manuscripts: Vol. 2: The 'Canterbury Tales'. Aldershot: Scolar Press.
Stolcke, A. (2002). SRILM: An extensible language modeling toolkit. In J. Hansen & B. Pellom (Eds.), Proceedings of the 7th International Conference on Spoken Language Processing (pp. 901–904). Denver: Casual Productions.
Stubbs, E. (Ed.) (2000). The Hengwrt Chaucer digital facsimile. Leicester: Scholarly Digital Editions.
Stubbs, E. (2007). 'Here's one I prepared earlier': The work of Scribe D on Oxford, Corpus Christi College, MS 198. Review of English Studies 58, 133–153.
Thaisen, J. (2008a). The Trinity Gower D Scribe's two Canterbury Tales manuscripts revisited. In M. Connolly & L. Mooney (Eds.), Design and distribution of late medieval manuscripts in England (pp. 41–60) York: York Medieval Press.
Thaisen, J. (2008b). Overlooked variants in the orthography of British Library, Additional MS 35,286. Journal of the Early Book Society for the Study of Manuscripts and Printing History, 11, 121–143.
Thaisen, J. (2009). Statistical comparison of Middle English texts: An interim report. Kwartalnik Neofilologiczny, 56(3), 205–221.
Thaisen, J. (in press). A probabilistic analysis of a Middle English text. In B. Nelson & M. Terras (Eds.), Digitizing medieval and early modern material culture (pp. 171–200). Tempe: Arizona Center for Medieval and Renaissance Studies. Also available from Iter: Gateway to the Middle Ages and Renaissance at http://www.itergateway.org.
Thaisen, J. & Da Rold, O. (2009). The linguistic stratification in the Cambridge University Library Dd copy of Chaucer's Canterbury Tales, Neuphilologische Mitteilungen 110, 295–309.
Thomson, C. (1998). A transcription and study of British Library MS. Lansdowne 851 of Chaucer's Canterbury Tales. Unpublished Ph.D. dissertation, University of Sheffield.
Vázquez, N. (2009). The 'Tale of Gamelyn' of 'The Canterbury Tales': An annotated edition. Lampeter: Edwin Mellen Press.
Wiggins, A. (2004). Are Auchinleck manuscript scribes 1 and 6 the same scribe?: Whole-data analysis and the advantages of electronic texts. Medium Aevum, 73(1), 10–26.
Witten, I. & Bell, T. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Transactions on Information Theory 37(4), 1085–1094.